Biochemical Properties of Random de Bruijn Sequences

Heitsch, Christine*, Li, Ming, Corn, Rob
University of Wisconsin at Madison, Madison, WI, USA

In the nucleus, lengthy DNA molecules have a canonical double-stranded helix structure well-adapted for information storage and retrieval. In the laboratory, short single-stranded DNA sequences have crucial biomedical applications, most notably in the extensive use of microarrays to diagnose genetic disorders. The contribution of microarray data to the prevention and treatment of diseases places the computational design and analysis of short DNA probe sequences at the forefront of current biological challenges.

Computer technology operates on a binary code of zeros and ones, however the genetic code is a four letter alphabet with energetics driven by the Watson-Crick base pairings. Thus, the essential challenge is designing sets of short oligonucleotides, or "DNA words," whose elements are strongly differentiated from each other *with respect to the biochemical energetics*. Under a solution to the DNA word design problem, each DNA segment W_i should bind strongly to its complement C_i. Figure 1 illustrates this binding on a surface array with a fluorescently labeled complement C_1. Furthermore, for $j \neq i$, there should be no significant binding between W_i and C_j (the complement problem), between W_i and W_j or C_i and C_j (the reverse complement problem), or of W_i with itself (the inverted repeat problem). The fundamental difficulty is addressing all three problematic interactions with a biologically relevant solution.

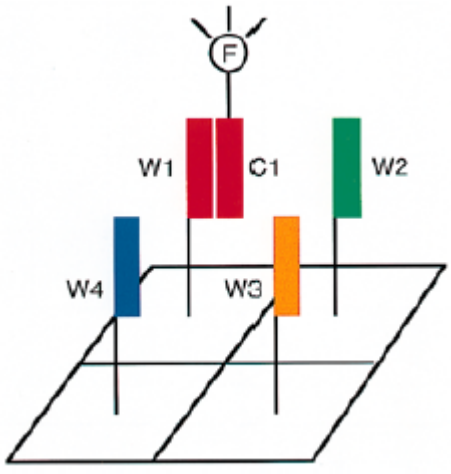

Figure 1: Microarray schematic

We present a solution to the DNA word design problem based on the biochemical properties of random de Bruijn sequences. De Bruijn sequences are part of a mathematical theory of strings, codes, and information which provides a foundation for addressing this computational biology question. Under the model that stable (mis)hybridization begins at a small region with perfect pairing, we control problematic interactions by preventing a nucleation complex. The complement problem is addressed by restricting repeated substrings with the adoption of de Bruijn sequences as our mathematical basis for noninteracting DNA segments. We provide an algorithm for generating these sequences uniformly at random from the 1.89×10^{20} total possibilities and analyze its performance. The program's output is first selected according to our criteria and then tested against the predicted biochemical properties. This solves the reverse complement and inverted repeat problems. We then experimentally verify the desired biochemical properties of our DNA words. Finally, we discuss our ability to engineer strings of nucleotide bases with specified characteristics as it pertains to current and future biomedical applications.

Christine Heitsch is supported by NLM training grant T15 LM07359, "Computation and Informatics in Biology and Medicine." This work with Prof. Rob Corn and Ming Li from the Corn research group in the Department of Chemistry is supported by grants from DARPA and NSF.